# A Heuristic Approach for Community Detection in Protein Networks

Sminu Izudheen[1] and Sheena Mathew[2]

[1] Department of Computer Science,
Rajagiri School of Engineering & Technology
sminu_i@rajagirtech.ac.in
[2] Division of Computer Science, School of Engineering,
Cochin University of Science & Technology
sheenamathew@cusat.ac.in

**Abstract.** Protein-protein interactions play a vital role in identifying the outcome of a vast majority of cellular mechanisms. But analyzing these complex data to identify community structures which can explain the activities of protein networks were always been a challenge. This paper reports the use of triangular modularity of protein network as an effective method to identify these community structures.

**Keywords:** Community Detection, Protein-Protein Interaction, Protein Networks, Spectral Optimization.

## 1 Introduction

Proteins involved in the same cellular processes often interact with each other, and these protein-protein interactions are fundamental to almost all biological processes [1]. The protein systems undergoing interactions with other polypeptides are particularly rich of natively unfolded tracts and these unfolded patches were discovered to be involved in both protein-protein interactions and aggregation in many different systems [2], [3]. Several efforts have been made to identify these interactions, so that biological systems can be understood better. With the emergence of a variety of techniques like yeast-two-hybrid, mass spectrometry and protein chip technologies, enormous amount of protein-protein interaction data are available [4]. However, due to the limitations in the techniques to handle such data, analysis of data in terms of biological function has not kept pace with data acquisition.

Protein complexes performing a specific biological function often contain highly connected protein modules [4]. These connected modules can be considered as community structures of protein networks. Even though community structures can better explain the activities of protein networks, this area is not well explored. But as identifying these community structures could be able to produce some useful findings, there exists some scope in investigating more on this. This motivated us to carry out

N. Meghanathan et al. (Eds.): CCSIT 2012, Part II, LNICST 85, pp. 469–475, 2012.

the present investigations on community structures and the results obtained prove that it is a promising method to detect community structures in protein networks.

A number of methods are proposed to detect community structures in complex networks. These include hierarchical clustering, graph partitioning based on network modularity, k-clique percolation, and many others [5]. Nevertheless, we preferred to make use of the decomposition algorithm (GN algorithm) proposed by Newman and Girvan due to its ability not only to divide networks effectively, but also to refuse to divide them when no good division exists.

## 2 Methods

### 2.1 Triangular Modularity Detection

The concept of community structure in complex networks was first pointed out in the by Girvan and Newman [6], and it refers to the fact that nodes in many real networks appear to group in subgraphs in which the density of internal connections is larger than the connections with the rest of nodes in the network. One of the most successful approaches to identify the community structure of complex networks is through the quality function called *modularity* [6], [7], which will define modules as well as provide a quantitative measure to find them. Here, we use motifs in the network to detect sub structures in a network. The modularity for weighted directed networks [8] is calculated as:

$$Q(C) = \frac{1}{2w}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(w_{ij} - \frac{w_i^{\text{out}} w_j^{\text{in}}}{2w}\right)\delta(C_i, C_j)\,, \tag{1}$$

where $w_{ij}$ is the weight of the connection from the i$^{\text{th}}$ to the j$^{\text{th}}$ node, $w_i^{out}=\sum_j w_{ij}$ represents the output strength and $w_j^{in}=\sum_i w_{ij}$ the input strength, $2w=\sum_{ij} w_{ij}$ gives the total strength of the network, $C_i$ is the index of the community to which the node i belongs, and the Kronecker $\delta$ is 1 if nodes i and j are in the same community, and 0 otherwise. For undirected networks, $w_i^{out}= w_i^{in} \equiv w_i$.

Given a partition C of an unweighted network, motif modularity can be represented as the fraction of motifs inside the communities minus the fraction in a random network [9], given by

$$Q_M(C) = \frac{\sum\limits_{i_1 i_2 \ldots i_M}\prod\limits_{(a,b)\in E_M} w_{i_a i_b}(C)}{\sum\limits_{i_1 i_2 \ldots i_M}^{n}\prod\limits_{(a,b)\in E_M} w_{i_a i_b}} - \frac{\sum\limits_{i_1 i_2 \ldots i_M}\prod\limits_{(a,b)\in E_M} n_{i_a i_b}(C)}{\sum\limits_{i_1 i_2 \ldots i_M}^{n}\prod\limits_{(a,b)\in E_M} n_{i_a i_b}}\,. \tag{2}$$

where

$$n^{ij} \quad = w_i^{out}\, w_j^{in},$$
$$w_{ij}(C) = w_{ij}\delta(C_i, C_j)$$
$$n_{ij}(C) = n_{ij}\delta(C_i, C_j)$$

This can be extended to find the community of triangles in a network. Applying equation (2), triangle motif can be expressed as

$$Q_{\triangle}(C) = \frac{\sum_{ijk} w_{ij}(C)w_{jk}(C)w_{ki}(C)}{\sum_{ijk} w_{ij}w_{jk}w_{ki}} - \frac{\sum_{ijk} n_{ij}(C)n_{jk}(C)n_{ki}(C)}{\sum_{ijk} n_{ij}n_{jk}n_{ki}} \, . \tag{3}$$

## 2.2 Triangular Modularity in Protein Networks

Protein interactions can be compared with an undirected graph with proteins as vertices and interactions as edges. We represented this interaction as an adjacency matrix. To detect the community structure in the protein network, we identified the triangular motifs in the network. Since we are considering an undirected graph, the triangle modularity [10] can be represented as

$$Q_{\triangle}(C) = \sum_i \sum_j \sum_k B_{ijk}\delta(C_i, C_j)\delta(C_j, C_k)\delta(C_k, C_i) \, . \tag{4}$$

where

$$B_{ijk} = \frac{w_{ij}w_{jk}w_{ki}}{\sum_i \sum_j \sum_k w_{ij}w_{jk}w_{ki}} - \frac{(w_i w_j)(w_j w_k)(w_k w_i)}{\sum_i \sum_j \sum_k (w_i w_j)(w_j w_k)(w_k w_i)} \, .$$

## 2.3 Spectral Optimization of Triangular Modularity in Protein Networks

Once we have the triangular modularity, next task is to define some optimization algorithm to calculate the modularity value. This is important since large number of traids can be formed. Here we propose spectral optimization [10] to perform this task. To detect the community structure in a network, eigen spectrum of the modularity matrix is used. We compute the leading eigenvector of the modularity matrix and divide the vertices into two groups according to the signs of the elements in this vector, with vertices whose corresponding elements are positive moves to one group and the rest moves to the other group. This process is repeated recursively, giving two partitions in each step until no new splits are possible.

One of the advantages of this algorithm over conventional partitioning methods is that, there is no need to constrain the group sizes or artificially forbid the trivial solution with all vertices in a single group. If there is no positive eigenvalues of the modularity matrix, then the leading eigenvector is the vector (1,1,1, …) corresponding to all vertices in a single group. In this case, the algorithm is telling us that there is no division of the network that results in positive modularity. Hence the algorithm has the ability not only to divide networks effectively, but also to refuse to divide them when no good division possible.

To perform spectral optimization on the modularity value calculated in equ.(4), we need to perform some transformations. In Belkacem Serrour et al [11], triangular modularity is reduced to standard spectral form as:

$$Q_{\triangle}(S) = \frac{3}{4} \sum_i \sum_j s_i M_{ij} s_j \ . \tag{5}$$

where

$$M_{ij} = \sum_k B_{ijk} \ .$$

## 2.4 Kernighan-Lin Optimization on Protein Networks

During each iteration of the algorithm, before dividing the network into two communities, the groups created by the spectral optimization is further improved by applying Kernighan-Lin optimization [12]. KL algorithm moves the vertices among the two groups to increase the modularity. For an arbitrary two-way partition S, the algorithm partition S into two sets A and B such that external cost is minimized. Suppose $A^*$ and $B^*$ represents a minimum cost two-way partition, then algorithm identifies $X \subset A$ and $Y \subset B$ with $|X| = |Y| \leq n/2$, such that interchanging X and Y produces A* and B*. In order to find X and Y from A and B without finding all possible choices, maximize the gain value,

$$g = D_a + D_b - 2c_{a,b} \ . \tag{6}$$

where, $c_{a,b}$ is the cost between vertices a and b, and $D_a$ and $D_b$ are the difference between external and internal cost given by

$$D_a = \sum_{y \in B} C_{ay} - \sum_{x \in A} C_{ax} \ . \tag{7}$$

$$D_b = \sum_{y \in A} C_{by} - \sum_{x \in B} C_{bx} \ . \tag{8}$$

We have applied KL optimization on sub groups created from Spectral optimization. Following steps are performed to identify *X* from *A*. First, *D* values for all elements in the group are calculated and the one with maximum value is selected as $a_1$. Second, set aside $a_1$ and recalculate *D* for the set $A$-$\{a_1\}$.Continue the same until all nodes are exhausted, identifying $a_1, a_2, .. a_n$. Repeat the same on *B* to identify $b_1, b_2, .. b_n$.

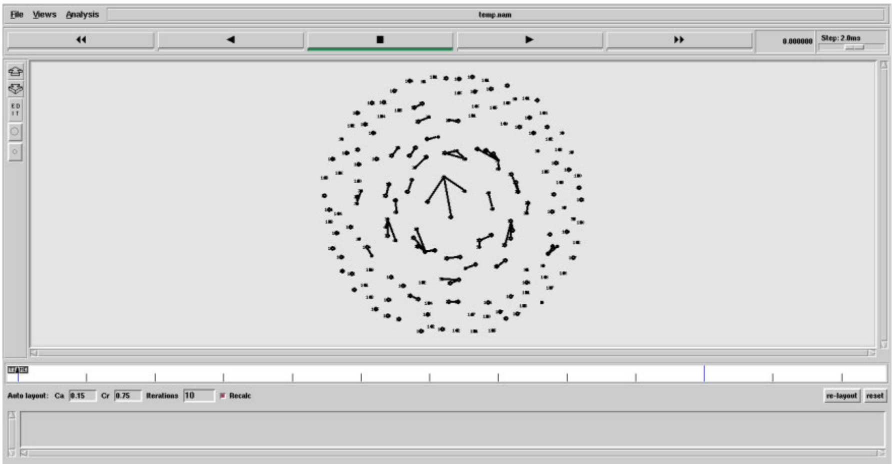
Calculate the corresponding gain $g_1, g_2, .. g_n$. Choose $k$ to maximize the gain $G=\sum_{i=1}^{k} g_i$ and select $X$ as $a_1, a_2, .. a_k$ and $Y$ as $b_1, b_2, .. b_k$. If $G$>0, reduction in cost of $G$ can be achieved, which means we can interchange $X$ and $Y$ between $A$ and $B$. If $G$=0, we have arrived at a local minimum and we have to repeat the steps by taking $a_2$ and $b_2$ as pivot elements. Results obtained shows that we will be able to reach a global optimum maximum in three iterations.

### 2.5 Dataset

For the present study protein interaction data is downloaded from MIPS [13] and MINT [14] databases.

## 3 Results

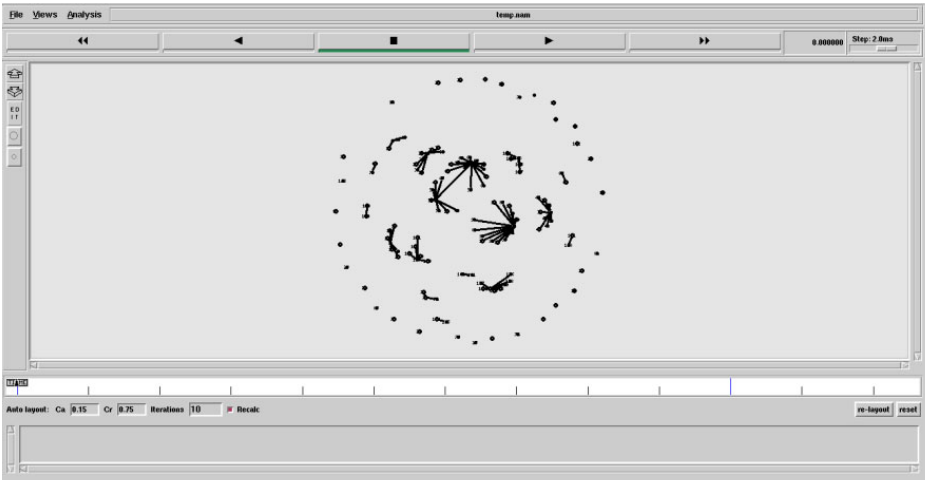
In this section we presents the results of the spectral optimization of triangular modularity applied to real protein interaction data from individually performed experiments. The results are then simulated using NS2. Fig.1 shows the simulated result for data downloaded from MIPS database. It represents the interaction between 193 different proteins represented as a protein interaction network. Here, shown as groups are the communities detected when we optimize the triangular modularity of the network.



Auto layout: Ca 0.15 Cr 0.75 Iterations 10 Recalc

**Fig. 1.** Communities detected by optimizing triangular modularity of protein interaction data downloaded from MIPS

Fig.2 represents the protein interaction network of 205 different human proteins downloaded from MINT. From the figure it is clear that the algorithm is able to detect communities from this network also.



Auto layout: Ca 0.15 Cr 0.75 Iterations 10 ☒ Recalc

**Fig. 2.** Communities detected by optimizing triangular modularity of protein interaction data downloaded from MINT

# 4 Conclusion

In this paper, we have demonstrated the use of triangular modularity as a promising method to analyze protein interactions. The method proved to be powerful in extracting community structures from protein networks. To show this, we have used Newman-Girvan algorithm to calculate triangular modularity. The modified algorithm have been tested on protein interaction data retrieved from databases like MIPS and MINT and are able to recover community patterns in protein networks. Hence community structure prediction proposed here can be applied to complex disease network to explore the relationship between human genetic disorders and the corresponding disease genes.

# References

1. Hakes, L., Lovell, S.C., Oliver, S.G., et al.: Specificity in protein interactions and its relationship with sequence diversity and coevolution. PNAS 104(19), 7999–8004 (2007)
2. Uversky, V.N., Segel, D.J., Doniach, S., Fink, A.L.: Association-induced folding of globular proteins. Proc. Natl. Acad. Sci. USA, 5480–5483 (1998,95)
3. Zbilut, J.P., Colosimo, A., Conti, F., Colafranceschi, M., Manetti, C., Valerio, M., Webber Jr., C.L., Giuliani, A.: Protein aggregation/folding: the role of deterministic singularities of sequence hydrophobicity as determined by nonlinear signal analysis of acylphosphatase and Abeta (1-40). Biophys. J. 85, 3544–3557 (2003)

4. Harwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W.: From molecular to modular cell biology. Nature 402, c47–c52 (1999)
5. Bader, D., et al.: Approximating betweenness centrality. Georgia Institute of Technology (2007)
6. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)
7. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113 (2004)
8. Arenas, A., Duch, J., Fernández, A., Gómez, S.: Size reduction of complex networks preserving modularity. New J. Phys. 9, 176 (2007)
9. Arenas, A., Fernández, A., Fortunato, S., Gómez, S.: Motif-based communities in complex networks. Journal of Physics A: Mathematical and Theoretical 41, 224001 (2008)
10. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences USA (103), 8577 (2006)
11. Serrour, B., Arenas, A., Gómez, S.: Detecting communities of triangles in complex networks using spectral optimization. Computer Communications (May 11, 2010)
12. Kernighan, B.W., Lin, S.: An efficient heuristic algorithm for partitioning graphs. The Bell System Technical Journal (February 1970)
13. Munich Information Center for Protein Sequences, `http://www.helmholtz-muenchen.de/en/mips/home/index.html`
14. MINT: the Molecular INTeraction database, `http://mint.bio.uniroma2.it/mint/Welcome.do`